UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ENHANCEMENT OF TEXT REPRESENTATION FOR INDONESIAN DOCUMENT SUMMARIZATION WITH DEEP SEQUENTIAL PATTERN MINING

DIAN SA'ADILLAH MAYLAWATI

DOCTOR OF PHILOSOPHY

2023

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

ENHANCEMENT OF TEXT REPRESENTATION FOR INDONESIAN DOCUMENT SUMMARIZATION WITH DEEP SEQUENTIAL PATTERN MINING

DIAN SA'ADILLAH MAYLAWATI

Doctor of Philosophy

2023

ENHANCEMENT OF TEXT REPRESENTATION FOR INDONESIAN DOCUMENT SUMMARIZATION WITH DEEP SEQUENTIAL PATTERN MINING

DIAN SA'ADILLAH MAYLAWATI

A thesis submitted
in fulfillment of the requirements for the degree of
Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I declare that this thesis entitled “Enhancement of Text Representation for Indonesian Document Summarization with Deep Sequential Pattern Mining” is the result of my own research except as cited in the references. The dissertation has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature :

Name : Dian Sa’adillah Maylawati

Date : April 06, 2023

APPROVAL

I hereby declare that I have read his thesis and in my opinion, this thesis is sufficient in terms of scope and quality for the award of the degree of Doctor of Philosophy.

Signature : ....

Supervisor Name : Dr. Yogan Jaya Kumar

Date : April 06, 2023

DEDICATION

I dedicate this thesis for:

My late father, who taught me hard work

My mother, who always prayed and instilled independence

My husband, who always supports and motivates

My children, who always give me joy

My brothers and sisters, who always inspire

My teachers, who educated me to achieve this dream

My best friends, who always color my life

ABSTRACT

Readability is a great challenge necessary to solve in text summarization research. Referring to the previous research studies, one key concern is minimizing the gap between the summary result and reader understanding. It is important to keep the meaning of the text to reach a readable summary result. However, every language has its grammar and structure characteristics. This also happens to the Indonesia language, in which a specific treatment is needed to find the meaning of the text. The present study hypothesizes that readability can be achieved with text representation that maintains the meaning of text documents well. Therefore, the present study aims: (1) to improve Indonesian text summary by enhancing the Sequence of Word (SoW) as text representation using Sequential Pattern Mining (SPM) with PrefixSpan algorithm since the effectiveness of SPM in Indonesian is proven useful for text classification and clustering; (2) to combine SPM and Deep Learning (DeepSPM) in text summarization with Indonesian text, as a result of its superior accuracy when trained with large amounts of data; and (3) to evaluate the readability of Indonesian text summary with several evaluation scenarios. Most text summarization research mainly uses co-selection-based analysis to evaluate the summary result. This seems to be less sufficient to evaluate readability. Therefore, this study includes content-based analysis and human readability evaluation to evaluate the readability of summary result. First, this study combines SPM with Sentence Scoring method as feature-based approach and Bellman-Ford algorithm as graph-based to validate the performance of SPM. Second, the proposed SPM approach is combined with Deep Belief Network (DBN), called DeepSPM, based on the unsupervised Deep Learning method. Then, the performance of the proposed methods in producing Indonesian text summary result is evaluated by Recall-Oriented Understudy for Gisting Evaluation (ROUGE) as co-selection-based analysis; Dwiyanto Djoko Pranowo metrics, Gunning Fog Index (GFI) and Flesch-Kincaid Grade Level (FKGL) as content-based analysis; and human readability evaluation. The experimental findings from this study, using IndoSum dataset, show that SPM can enhance the quality of summary results. DeepSPM achieves better results than DBN with f-measure scores of 46.21% for ROUGE-1, 36.94% for ROUGE-2, and 41.01% for ROUGE-L. Furthermore, the readability evaluation using Dwiyanto's metrics, GFI, and FKGL also shows that the summary results of DeepSPM are readable at a moderate level and are consistent with the human evaluation results conducted by two Indonesian language experts.

PENINGKATAN PERWAKILAN TEKS UNTUK RINGKASAN DOKUMEN INDONESIA DENGAN PERLOMBONGAN CORAK TURUTAN MENDALAM

ABSTRAK

Kebolehbacaan ialah cabaran hebat yang perlu diselesaikan dalam penyelidikan ringkasan teks. Merujuk kepada kajian penyelidikan terdahulu, salah satu kebimbangan utama adalah untuk meminimumkan jurang antara hasil rumusan dan pemahaman pembaca. Adalah penting untuk mengekalkan maksud teks untuk mencapai hasil ringkasan yang boleh dibaca. Walau bagaimanapun, setiap bahasa mempunyai ciri-ciri tatabahasa dan strukturnya sendiri. Ini juga berlaku kepada bahasa Indonesia di mana rawatan khusus diperlukan untuk mencari makna teks. Kajian ini membuat hipotesis bahawa kebolehbacaan boleh dicapai dengan perwakilan teks yang mengekalkan makna dokumen teks dengan baik. Oleh itu, kajian ini bertujuan: (1) untuk menambah baik ringkasan teks bahasa Indonesia dengan meningkatkan Sequence of Word (SoW) sebagai representasi teks menggunakan Sequential Pattern Mining (SPM) dengan algoritma PrefixSpan memandangkan keberkesanan SPM dalam bahasa Indonesia terbukti berguna untuk klasifikasi teks. dan pengelompokan; (2) untuk menggabungkan SPM dan Pembelajaran Dalam (DeepSPM) dalam ringkasan teks dengan teks Indonesia, hasil ketepatannya yang unggul apabila dilatih dengan jumlah data yang besar; dan (3) untuk menilai kebolehbacaan ringkasan teks bahasa Indonesia dengan beberapa senario penilaian. Kebanyakan penyelidikan ringkasan teks terutamanya menggunakan analisis berasaskan pemilihan bersama untuk menilai hasil rumusan. Ini nampaknya kurang mencukupi untuk menilai kebolehbacaan. Oleh itu, kajian ini merangkumi analisis berasaskan kandungan dan penilaian kebolehbacaan manusia untuk menilai kebolehbacaan hasil rumusan. Pertama, kajian ini menggabungkan SPM dengan kaedah Penskoran Ayat sebagai pendekatan berasaskan ciri dan algoritma Bellman-Ford sebagai berasaskan graf untuk mengesahkan prestasi SPM. Kedua, pendekatan SPM yang dicadangkan digabungkan lagi dengan Deep Belief Network (DBN), dipanggil DeepSPM yang berasaskan kaedah Pembelajaran Dalam tanpa pengawasan. Kemudian, prestasi kaedah yang dicadangkan dalam menghasilkan hasil ringkasan teks bahasa Indonesia dinilai oleh Recall-Oriented Understudy for Gisting Evaluation (ROUGE) sebagai analisis berasaskan pemilihan bersama; Metrik Dwiyanto Djoko Pranowo, Gunning Fog Index (GFI) dan Tahap Gred Flesch-Kincaid (FKGL) sebagai analisis berasaskan kandungan; dan penilaian kebolehbacaan manusia. Penemuan eksperimen daripada penyelidikan ini, menggunakan dataset IndoSum, menunjukkan bahawa SPM boleh meningkatkan kualiti keputusan ringkasan. DeepSPM mencapai keputusan yang lebih baik daripada DBN dengan skor f-measure sebanyak 46.21% untuk ROUGE-1, 36.94% untuk ROUGE-2, dan 41.01% untuk ROUGE-L. Tambahan pula, penilaian kebolehbacaan menggunakan metrik Dwiyanto, GFI dan FKGL juga menunjukkan bahawa keputusan ringkasan DeepSPM boleh dibaca pada tahap sederhana dan konsisten dengan keputusan penilaian manusia yang dijalankan oleh dua pakar bahasa Indonesia.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Gracious, the Most Merciful

First and foremost, I would like to thank and praise Allah the Almighty, my Creator, my Sustainer, for everything I received since the beginning of my life. I would like to extend my appreciation to the Universitas Islam Negeri (UIN) Sunan Gunung Djati Bandung and Universiti Teknikal Malaysia Melaka (UTeM) for providing the research platform.

My utmost appreciation goes to my main supervisor, Dr. Yogan Jaya Kumar, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), for all his support, advice and inspiration. His constant patience for guiding and providing priceless insights will forever be remembered. Also, to my co-supervisor, Dr. Fauziah binti Kasmin, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), who constantly supported my journey. My special thanks go to Prof. Dr. H. Muhammad Ali Ramdhani, S.TP., MT as Professor of Information Technology UIN Sunan Gunung Djati Bandung and Dr. Basit Raza from Department of Computer Science, COMSATS University Islamabad (CUI) for all the help and support I received from them. Also thank you for Dr. Halizah binti Basiron and Dr. Nur Zareen binti Zulkarnain for the advices on every presentation of my dissertation progress.

Last but not least, from the bottom of my heart gratitude to my beloved husband, Ayah Gunawan, for his encouragement and who has been the pillar of strength in all my endeavors. My eternal love also to all my children, Adhaya Desela Diguna and Shafa Ayudisa Diguna, for their joy, patience and understanding. I would also like to thank my beloved parents, my late father Bapak Mursyidi and mother Mamah Siah Khosyi'ah, for their endless support, love and prayers. Finally, thank you to all the individual(s) (brother Royhan Aziz Ghifari, sister Anggi Rufaedah Wardani, friends at Department of Informatics UIN Sunan Gunung Djati Bandung, friends at Academic Writing Class, friends at Pascasarjana UIN Sunan Gunung Djati Bandung, friends in taking a Ph.D. in UTeM, my students) who had provided me the assistance, support and inspiration to embark and finish on my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

5W1H	-	Why, Who, Where, When, What and How
AE	-	Auto-Encoder
AI	-	Artificial Intelligence
AKNN	-	Adaptive K-Nearest Neighbor
ANN	-	Artificial Neural Network
ARI	-	Automated Readability Index
BART	-	Bidirectional Auto-Regressive Transformers
BERT	-	Bidirectional Encoder Representations from Transformers
BIDE	-	Bi-Directional Extention
Bi-LSTM	-	Bidirectional Long Short-Term Memory
BoW	-	Bag of Words
CD	-	Contrastive Divergence
CNN	-	Convolutional Neural Network
CLI	-	Coleman-Liau Index
CPLNVN	-	Centroid, Position, Sentence Length, Noun Verb and Numerical Data
CRF	-	Conditional Random Fields
DeepSPM	-	Deep Sequential Pattern Mining
DAG LSTM	-	Directed Acyclic Graph Long Short-Term Memory
DIMASP-C	-	Discover all the Maximal Sequential Patterns (document collection)
DIMASP-D	-	Discover all the Maximal Sequential Patterns (single document)
DQN	-	Deep Q-Networks
DBN	-	Deep Belief Network
EDA	-	Exploratory Data Analysis
ENAE	-	Ensable Noisy Auto-Encoder
FASP	-	Frequent Adjacent Sequential Pattern
FASPe	-	Frequent Eliminated Pattern
FKGL	-	Flesch-Kincaid Grade Level

FPM	-	Frequent Pattern Mining
FWI	-	Frequent Word Itemsets
FWS	-	Frequent Word Sequence
GA	-	Genetic Algorithm
GFI	-	Gunning Fog Index
HAN	-	Hierarchical Attention Network
IEEE	-	Institute of Electrical and Electronics Engineers
IndoBART	-	Indonesian based on the BART model
IndoBERT	-	Indonesian version of BERT model
IndoLEM	-	Indonesian Language Evaluation Montage
IndoNLG	-	Indonesian Language for Natural Language Generation
IndoNLU	-	Indonesian Natural Language Understanding
IndoSum	-	Indonesian Summarization
KCSP	-	Key-phrase Candidate Search using sequential Pattern
LCS	-	Long Common Subsequence
LIX	-	The Lasbarhetsindex Swedish Readability Formula
LDA	-	Latent Dirichlet allocation
LSTM	-	Long Short-Term Memory
LSTM-CRF	-	Long Short-Term Memory Conditional Random Field
MCBA	-	Modified Corpus-based Approach
MFS	-	Maximal Frequent Sequence
MoW	-	Multiple Words
MWI-Sum	-	Multilingual Weighted Itemsetbased Summarizer
NAE	-	Noisy Auto-Encoder
NATS	-	Neural Abstractive Text Summarizer
NER	-	Named Entity Recognition
NLP	-	Natural Language Processing
NLTK	-	Natural Language Toolkit
NR	-	Normal Ratio
OVIX	-	Ordvariationsindex, Swedish word variation index
POS	-	Part of Speech

PRISMA	-	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RBM	-	Restricted Boltzmann Machine
RNN	-	Recurrent Neural Networks
ROUGE	-	Recall-Oriented Understudy for Gisting Evaluation
RPS	-	Relatedness with Previous Sentence
S-LSTM	-	Sentence state Long Short-Term Memory
SFWI	-	Set of Frequent Word Itemset
SFWS	-	Set of Frequent Word Sequence
SINTA	-	Science and Technology Index
SLR	-	Systematic Literature Review
SMOGI	-	Simple Measure of Gobbledygook Index
SML	-	Supervised Machine Learning
SoW	-	Sequence of Words
SPM	-	Sequential Pattern Mining
SPMW	-	Sequential Pattern Mining Wildcard
SWING	-	Summarizer from the Web Information Retrieval/ NLP Group
TAC	-	Text Analysis Conference
TF	-	Term Frequency
TF-IDF	-	Term Frequency – Inverse Document Frequency
TF-ISF	-	Term Frequency – Inverse Sentence Frequency
VSM	-	Vector Space Model

LIST OF SYMBOLS

∂	-	Derivative value
η	-	Learning rate value
Syl	-	Syllable
tf	-	Term frequency
idf	-	Inverse document frequency
isf	-	Inverse sentence frequency
∞	-	Infinity value
θ	-	Parameter of probability function
β	-	Positive real factor, precision devided by recall

LIST OF PUBLICATIONS

1. Maylawati, D. S., Kumar, Y. J., Kasmin, F. B., and Ramdhani, M. A., 2019. An idea based on sequential pattern mining and Deep Learning for text summarization. In *Journal of Physics: Conference Series* (Vol. 1402, No. 7, p. 077013). IOP Publishing. (SCOPUS indexed)

2. Maylawati, D. S. A., Kumar, Y. J., Kasmin, F. B., and Raza, B., 2019. Sequential pattern mining and Deep Learning to enhance readability of indonesian text summarization. *International Journal of Advanced Trends in Computer Science and Engineering. https://doi. org/10.30534/ijatcse/2019/78862019*. (SCOPUS indexed)

CHAPTER 1

INTRODUCTION

1.1 Introduction

Most people use the internet and digital technology in the era of big data. Almost all sectors utilize technology, such as the education sector, which utilizes digital platforms for learning, the economic sector which gives rise to various e-commerce, the banking sector which utilizes digital and cashless transactions, to socialization which utilizes social media, blogs, forums and news portals. These digital activities contribute data with big volume, rapid velocity flow, variety of types, important value content, and veracity certainty. One of these abundant data sources is text data from various news sites, blogs, and social media. Natural Language Processing (NLP) is a technology used mainly for text data.

NLP technology is developed rapidly in the era of big data today. NLP, also based on artificial intelligence technology, can process many types of languages, either in the lexical, semantic, or syntactic (Nadkarni et al., 2011); (Pandey and Rajput, 2020). One popular NLP application is automatic text summarization, which produces summaries from document collections using either an extraction or abstraction approach (Yulyardo et al., 2018). A summary is a condensed version of a document's content that includes most of the information found in the original text(s) (Hovy and Marcu, 2005). Text summarization involves content reduction and generalization based on what is relevant in the source text to produce the summary. The process of automatically constructing such summaries, using a computer is known as automatic text summarization. Automatic text summarization research frequently discovers new techniques to construct summaries to meet the needs of different applications and users.

However, every language is unique. Each language has its own grammar structures and rules, including the Indonesian language. Indonesian grammar is divided into two parts: morphology and syntax (Alwi et al., 2003); (Sneddon, 2003); (Tim Pengembang Pedoman Bahasa Indonesia, 2016). Morphology discusses the grammatical structure of the Indonesian language such as: absorption words, affixes (prefixes, suffixes, infixes), and so on. They are related to the origin of word-formation. Meanwhile, syntax is broader than morphology. It is related to sentences, relationships between words, and deals with grammar within speech. In other words, it holds the speech's meaning, content, purpose, or ideas.

Many NLP communities in Indonesia prepare Indonesian datasets and conduct NLP research, including text summarization. Several NLP communities in Indonesia build Indonesian NLP benchmarks, such as: Indonesian Summarization (IndoSum) (Kurniawan and Louvan, 2018), Indonesian Language Evaluation Montage (IndoLEM) (Koto et al., 2020), Indonesian Natural Language Understanding (IndoNLU) (Wilie et al., 2020) and Indonesian Language for Natural Language Generation (IndoNLG) (Cahyawijaya et al., 2021). IndoLEM is developed from IndoSum, conducts a better method, and provides more news datasets for Indonesian text summarization research.

Many methods can be used for automatic text summarization. The basic method is the Sentence Scoring method (Sri et al., 2017); (Sabuna and Setyohadi, 2017), Graph-Based method (Garmastewira and Khodra, 2019), Machine Learning (Patel et al., 2018) and Deep Learning (Padmapriya and Duraiswamy, 2014); (Yousefi-Azar and Hamey, 2017); (Adelia et al., 2019). IndoSum and IndoLEM also use Deep Learning. IndoSum uses Long Short-Term Memory (LSTM), while IndoLEM uses Bidirectional Encoder Representations from Transformers (BERT).